Available online at www.sciencedirect.com
ScienceDirect
Procedia - Social and Behavioral Sciences 97 (2013) 691 – 699
Procedia
Social and Behavioral Sciences

The 9th International Conference on Cognitive Science

Group usability testing of virtual reality-based learning environments: A modified approach

Chwen Jen Chen[a,*], Siew Yung Lau[a], Kee Man Chuah[b], Chee Siong Teh[a]

[a]*Faculty of Cognitive Sciences and Human Development, Universiti Malaysia Sarawak, Kota Samarahan, 94300, Malaysia*
[b]*Centre for Language Studies, Universiti Malaysia Sarawak, Kota Samarahan, 94300, Malaysia*

Abstract

Conventional usability testing is usually conducted with several individual participants. In recent years, however, group usability testing is gradually gaining attention. Such approach involves several-to-many participants performing tasks simultaneously, with one to several testers observing and interacting with the participants. This approach is able to generate many useful data within a short period of time. In light with the need to further improve the approach, this paper presents a modified version of a group usability testing and how it can be feasibly used to evaluate the usability of a non-immersive virtual reality-based learning environment. The proposed modified group approach aims to minimize the possibility of data loss during the usability testing process. The effectiveness and efficiency of this modified method was compared to the original approach of group usability testing. The results indicate that the modified group usability testing is more effective and efficient than the original approach as it can collect more critical and significant data with lesser time, cost and effort consumption.

Keywords: usability testing; group usability testing; non-immersive virtual reality

1. Introduction

Usability testing has been widely used as an important technique to uncover the possible usability problems of a system. Poor usability of a system could prevent its effectiveness and efficiency of use [1]. Usability testing is defined as "a process that employs participants who are representative of the target population to evaluate the degree to which a product meets specific usability criteria" [2]. Generally, usability testing involves a process of observing users while using systems, and thereby extracts the usability issues from these users. Usability testing also intends to obtain feedback from representative users of a system in order to identify usability problems. Such testing is important in discovering major usability problems that are caused by human error, which may lead to confusion or termination of interaction with the system as well as frustration.

Usability studies of virtual reality (VR), especially non-immersive VR, are still insufficient although VR is an advanced technology which is gaining widespread acceptance in various fields particularly in education [3, 4, 5]. A review of the literature has shown that some researchers have adopted traditional techniques in testing and evaluating non-immersive VR, or Desktop VR. Villanueva has evaluated the non-immersive VR, desktop, photo-

* Corresponding author. Tel.: +6082-581562; fax: +6082-581567
E-mail address: cjchen@fcs.unimas.my

doi:10.1016/j.sbspro.2013.10.289

realistic virtual environments using think-aloud protocol and heuristic evaluation [4]. Marsh and Wright [6] have carried out a usability test on non-immersive VR system using the co-operative evaluation while Rosli et al. [7] have conducted a usability testing on the interface of VR-based learning environment using the Software Usability Measurement Inventory (SUMI). In addition, Costalli et al. [8] have presented a set of criteria that should be considered in order to obtain a usable virtual environment. Rezazadeh, Firoozabadi and Wang [9], on the other hand, evaluated a virtual environment using affective measures in order to uncover usability issues.

It is rather apparent that limited usability studies are conducted on non-immersive VR. Studies which have been carried out on such VR systems, however, mostly involve several individual participants. There is still no approach which involves groups of participants to be tested simultaneously. Hence, this study looks into the potential and benefits of employing group approach in testing the usability of a non-immersive VR system. The approach presented in this paper is a modified group usability testing in which the original version is proposed by Downey [10]. This approach is adopted in the usability testing of a non-immersive VR-based system, which was developed and used as a case for the trial of the proposed approach. The effectiveness and efficiency of this modified group usability approach are examined and compared. In addition to that, the benefits and drawbacks of this approach are also discussed.

2. Group Usability Testing

Downey [10] defined group usability testing as an approach which involves several-to-many participants performing task at the same time. One to several testers are required to observe and interact with participants. Based on Downey's research, group usability testing is good at uncovering major usability issues. Downey also claims that with this approach, the criticality of the problems identified is able to be validated fast based on the frequency that the problem been pointed out by the participants.

Basically, group usability testing involves three stages; (1) user profile survey, (2) basic tasks exercises and (3) usability issues discussion. In this study, prior to the usability testing, representative tasks were selected. Before the testing was carried out, a brief training was conducted for the participants to provide them with some basic information on the tested system as well as their tasks. They were informed on the presence of observers to observe their interaction with the system.

Two identical cycles of group usability testing were conducted. Each cycle involved three stages, which were the user profile survey, basic task exercise, and usability issues discussion. However, there were no significant changes made to the system between the two cycles of testing. The following usability activities were carried out in temporal order.

2.1. User Profile Survey

In this first stage of the usability testing, Downey proposes a simple user profile survey on the participants in order to classify them based on some pre-determined characteristics. The purpose of this user profile survey is to ensure the homogeneity of participants who undergo the usability testing process later. This session may take 20 to 30 minutes depending on the number of participants. In this study, the target users of were lower secondary students. The user profile survey was obtained from the teachers of the chosen schools and via a simple survey form. The possible users' characteristics for classifications were gender, academic performance, computer literacy and English language proficiency. These characteristics were identified in order to choose participants who were qualified to involve in the usability testing. Gender and academic performance were used to ensure that the groups of participants recruited were almost equally distributed in terms of these characteristics. All participants were also required to have basic computer literacy skill to ensure their capability to use generic computer input devices, such as mouse and keyboard to interact with the system. The characteristic for English language proficiency was used to ensure the students were capable to understand the information presented by the system and to verbalize their opinions during the usability issues discussion session.

2.2. Basic Task Exercise

This session is the core of the group usability testing process. It takes approximately one hour, depending on the tasks given. Selected basic tasks are given to the participants using a set of written instructions. For the purpose of the study, the basic tasks given in this testing were to identify objects which might cause fire and to identify flammable objects.

The participants were seated in a setting which was somewhat circular although Downey [10] suggests a strictly circular setting. Such setting did not allow a participant to view the screens of the participants sitting next to them unless he/she purposefully leaning over to see those screens. The participants were allowed to ask questions during the testing. Participants performed the basic tasks given to them individually, but simultaneously. Meanwhile, multiple observers or testers, often the usability experts and/or software or system developers, walked around the circular setting and record usability issues faced by participants in both cycles of testing. Downey recruited three observers in her study. Therefore, in this study, three observers, comprised a system developer and two other individuals with sound usability knowledge, were involved.

During the testing session, the observers were allowed to occasionally interact with participants, answer questions, and also minimally probe them. Besides, discussion among the observers was also allowed as long as it did not cause disturbance to the participants. Such discussion facilitated the observers to prioritize the criticality of the recorded usability issues. Observers focused only on new usability issues during the second cycle of testing.

2.3. Usability Issue Discussion

After completing the basic tasks exercise session, Downey suggests the observers to facilitate a discussion on the usability issues or problems recorded by the observers and those raised by the participants. This session may also takes up one hour, depending on the amount and criticality of the problems identified. This discussion also facilitates the probing of clarification and perceived difficulties during the basic tasks exercises session and aims to prioritize the usability problems. In the subsequent text, Downey's group usability testing will be referred to as DGUT.

3. Modified Group Usability Testing

This study proposes a few modifications to the setting and procedure of DGUT, with the intention to enhance the effectiveness and efficiency of DGUT. This modified group usability testing will be referred to as MGUT. This section provides a description of the proposed modifications.

3.1 Basic Tasks Exercise

In the original approach proposed by Downey, only observers are required to do the recording of the issues or observations. This may result in the loss of some data as the observers who are responsible to record all the usability issues may not be able to observe and record different usability problems that are revealed or faced simultaneously by different participants.

In the proposed modified approach, besides having the observers to do the recording of their observations, participants are also asked to briefly jot down usability issues that they encounter during the testing session. Such arrangement is anticipated to be able to gather more usability data. Besides, in this modified approach, screen recording software is used to record the participants' interaction with the system during the usability testing session. Such recording enables observers to make a reference to it during the discussion session, particularly when confusion or discrepancies arise. Besides, the recorded video files of the interaction can also be used to assist in discovering trends and extracting more performance data when needed.

As for the setting of the computer configuration, MGUT allows a more flexible setting than DGUT as the computers in MGUT can just be grouped in a somehow circular setting while DGUT needs a strictly circular setting.

3.2 Usability Issues Discussion

In addition to what is done in the similar stage of DGUT, participants of MGUT are also prompted [11] by the observers to actively verbalize the usability issues or problems that they have jotted down during the testing session. This helps the participants to recall their interaction with the system and once again aims to minimise the possibility of losing any usability issues.

4. Methods of Data Analysis

4.1 VrSAFE

For the purpose of this study, VrSAFE, a non-immersive VR-based learning environment that educates its users on home fire safety and prevention, was developed and used as an application that was tested in this study. VrSAFE consists of a non-immersive three-dimensional (3D) virtual environment which is integrated onto a web interface that contains other multimedia elements, such as images, sound and animation. Figure 1 shows a screenshot of VrSAFE.

Fig. 1. A screenshot of VrSAFE

4.2 Usability Problems

Every identified usability problem was classified into its appropriate usability criterion/criteria, usability factor(s), scope and level of severity.

These problems were independently coded and categorized into different usability criteria based on the QUIM model proposed by Seffah et al. [12], which is a consolidated, hierarchical model of usability measurement that unifies various usability standards and conceptual models. Observers were responsible to discuss and reach consensus on the appropriate usability criterion/criteria for each usability problems. The usability criteria were then linked to their related usability factors based on the relationship table between usability criteria and usability factors as proposed in QUIM.

The scope of a usability problem refers to how narrowly or how widely the problem occurs [1]. The scope of a problem can be characterized as either local or global [13]. A local problem affects only one particular part of the system, while a global problem might affect more than one parts, therefore indicating the broad-based problem of the system that might affect the entire system.

The severity of a usability problem identified is the frequency, impact and persistence of the problem [14]. The severity of a problem is rated independently via a severity rating scale [13], which consists of four scales as follows:

- Level 1 -- problems that prevent completion of a task
- Level 2 -- problems that create significant delay and frustration

- Level 3 -- problems that have a minor effect on usability
- Level 4 -- problems that are more subtle and often point to an enhancement that can be added in the future

4.3 Comparison Technique

The comparison is based on the following questions:

Number of problems identified

- How many usability problems were identified by DGUT and MGUT respectively [15]?

Nature of problems identified

- How many problems were identified for each usability criterion by each testing approach?
- How many problems were identified for each scope by each testing approach?
- How many problems were identified for each severity level by testing approach?

Time/cost efficiency and participant richness

- What is the time taken to complete each testing approach?
- What is the number of participants involved in each testing approach?

4.4. Comparison Criteria

The list of criteria used to compare the different testing approaches is as follows:

- *Ability to detect problems* [16, 17]. The ability to detect problems refers to the number of problems identified by each usability testing approach.
- *Quality of problems identified* [16, 17]. The quality of problems identified refers to how well a usability testing approach in identifying useful and critical problems. Hence, it compares the usability factors, in the attempt to investigate the variedness of problems identified. The comparison of the scope and the severity of the problems also crucial to examine how well a usability testing approach in identifying major and significant problems.
- *Participant richness* [9, 18]. The number of participants involved in a usability testing approach indicates the participants' richness as there is power in number and therefore produces more convincing data and result.
- *Time, effort and cost-effectiveness* [15, 16, 17]. Time refers to the total time taken to carry out the usability testing approach while effort refers to the required work in carrying out the usability testing including the preparation work, such as user recruitment, computer setting and so forth. Cost-effectiveness refers to how well a usability testing approach is able to collect more and critical data with minimal time, cost and effort consumption.

5. Results

The modified approach (MGUT) is compared with the original version (DGUT) and the identified usability problems are tabulated accordingly. Table 1 shows a summary of the comparison.

Table 1. Comparison of DGUT and MGUT

Usability Problems identified (UPs)		DGUT	MGUT
Total UPs		30	43
Usability Factor	Effectiveness	4	10
	Efficiency	8	17
	Satisfaction	13	19
	Learnability	13	21
	Safety	1	0
	Universality	20	29
	Usefulness	7	15
Scope	Global UPs	18	30
	Local UPs	12	13
Severity	Level 1	3	3
	Level 2	5	13
	Level 3	16	19
	Level 4	6	8
Time taken (minutes)		120	105
No. of Participants involved		36	36

5.1. Ability to Detect Problems

The results indicate that MGUT managed to identify more usability problems, which are 43 as compared with DGUT, which identified 30 usability problems. The modifications employed in MGUT most probably explain the effectiveness of the approach in identifying usability problems. MGUT revealed 13 more usability problems than DGUT. In MGUT, users were required to briefly jot down the usability problems that they encountered during the testing process. This process of jotting down had enriched the usability discussion session as the notes served as a useful reference for the participants to recall the usability problems that they encountered while engaging in VrSAFE. Besides, in MGUT, the participants had more chance to elaborate and express their opinions as they were prompted by the observers via certain probing questions. This is also one of the modifications added to MGUT in order to discover more problems during the usability discussion session. It is worthy to note that when the participants are actively verbalizing in the usability discussion session, the issues raised by a participant often influenced other participants, which drove them to give their opinions and comments as well.

5.2. Quality of Problems Identified

The quality of problems identified refers to how well a usability testing approach in identifying useful and critical problems [16, 17]. As shown in Table 1, it is quite obvious that MGUT is better in revealing global usability problems. MGUT revealed 30 global usability problems while DGUT with 18 global usability problems. Global usability problems are more broad-based problems which will affect the system in more than one part [13]. The discussion among participants and observers, which revealed more problems with different perspective and view, could be a contribution towards the capability of MGUT to identify this type of problems.

The discussion sessions that were conducted after the individual interaction session with the system had naturally moved the focus of the participants from the specific aspect or feature of the system to the discussion of the more general context of the usability problems. MGUT can identify usability problems not only from observers' perspective, but also from users' perspective. The users were observed to discuss the usability problems from a more general perspective while observers tend to refer the problem to a specific component of the system. One of the examples of global usability problem identified by MGUT are *"Map was not sufficiently detail thus causing the users to be unsure of where the virtual objects were placed"*, *"Some users commented that the graphic /visual*

design of the system was unattractive", and *"Some users commented that the task given was too simple"*. These global usability problems were pointed out by participants during the usability discussion session in both MGUT and DGUT.

In terms of severity, both testing approaches are capable of identifying the most critical and serious usability problems, which are categorized as Level 1. The usability problems of this level are most severe as they prevent the completion of a task. Level 2 and Level 3 consist of moderate usability problems, where the usability problems of Level 2 are those problems which create significant delay and frustration while the usability problems of Level 3 are those problems which have a minor effect on usability. The results reveal that MGUT is more capable to record problems which fall into Levels 2 and 3. In DGUT, the observers are given the total responsibility to record usability problems. Therefore, they potentially missed many critical usability problems as they had to walk around and observe different individual participants. MGUT, the participants also took part in recording usability problems while exploring the system. This modification helped in overcoming data loss problems due to participants' forgetfulness and/or observers' less attentiveness.

5.3. Participant Richness

DGUT and MGUT which involved 36 participants are richer in user involvement as compared to other individual-based usability testing. Both DGUT and MGUT allow several-to-many participants to be simultaneously tested, in a relative short period of time. The results of DGUT and MGUT are more likely to convince others as these approaches involve more participants. This increase the reliability of the data collected which is important to convince the system developer to make changes to the system as a little change to the system might be costly. Besides, the usability issue discussion supports the trend toward richer user involvement in problem identification, collaborative design and evaluation experience.

5.4. Time, Effort and Cost-effectiveness

As indicated in Table 1, MGUT consumed slight lesser time than DGUT (a difference of about 15 minutes). From the empirical testing aspect, the preparation of the physical setting of DGUT and MGUT needs more efforts compared with an individual testing approach. However, such preparation work of MGUT is easier than DGUT. MGUT allows a more flexible computer arrangement setting than DGUT, which is confined to a circular setting. Hence, MGUT is more cost-effective than DGUT as it can collect more critical and significant data with lesser amount of time, cost and effort consumption.

6. Discussion

The most obvious strength of the proposed MGUT is the various steps taken to minimize data loss during the usability testing process. These steps include requiring the participants to be involved in the recording of the usability issues. They are required to jot down any usability problems that they encountered. This is crucial as the main purpose of usability testing is to gather as much as possible useful and critical data. This is an enhancement to the DGUT where observers are given the sole responsibility to record data. The observers potentially miss many usability issues as they just walk around and make observations of many different individuals.

The proposed use of screen recording software in the modified group approach also helps to produce a backup resource for the usability issues. With this means, there exists a source where the observers can refer to when there is a need in convincing the justification made to a usability problem.

The main reason that most of the usability testing approaches involve only few participants is because recruiting more users requires more time, effort and higher cost. Essentially, DGUT and MGUT allows several-to-many participants to be simultaneously tested, and produces lots of useful and critical usability data in a relative short period of time. This saves the observer's time and increases the cost-effectiveness of the usability testing process. In addition, MGUT also enables major usability issues to be revealed in a very short time. As pointed out by Downey [9], this testing approach can also take the advantage of the availability of many subjects who gather together in one place. Furthermore, the results of DGUT and MGUT are more likely to convince others as such group approach

considers more participants. The focused discussion at the final stage of the testing also supports the trend toward richer user involvement in problem identification, collaborative design and evaluation experience.

As pointed out by Downey [9], the most apparent drawback of group usability testing is that participants tend to affect one another during the testing session, especially during the discussion session. A participant may be influenced by other participants' opinions. Besides, this approach requires several observers, which may be limited in some cases. The availability of a computer lab that allows flexible arrangement of computers is also sometimes limited. As mentioned earlier, this approach is good at uncovering major usability problems, thus it is not competent in gathering more detailed problems as the interaction between the observers and participants is kept minimal.

7. Conclusion

This paper has presented a useful approach of usability evaluation of a non-immersive VR system. In general, it proposes an enhanced version of the group usability approach in order to improve the data gathering process by minimizing the possibility of data loss. The detail procedure for this modified group usability approach was explained. The effectiveness and efficiency of this modified group usability approach was examined via a comparison between this proposed approach with the original approach. The result shows that the modified group usability testing is more effective and efficient than the original approach as it can collect more critical and significant data with lesser time, cost and effort consumption. Lastly, the benefits as well as drawbacks of this modified group usability testing are discussed. The modified approach can also be applied to test other systems and not confined to VR-based learning environments as it contains sets of methods which can be universally implemented.

Acknowledgements

The author acknowledges the financial and facilities support rendered by Universiti Malaysia Sarawak to accomplish this project and disseminate the findings at this conference.

References

[1] Barnum, C.M. Usability testing and research, Pearson Education, New York, 2002.
[2] Rubin, J. Handbook of usability testing: How to plan, design, and conduct effective tests, Wiley, New York, 1994.
[3] Morar, S.S., Macredie, R.D. Special issue on "interacting with non-immersive virtual environments: Perception and navigation". Virtual Reality 7, 2004, pp.129-130.
[4] Villanueva, R., Moore, A., William Wong B.L. Usability evaluation of non-immersive, desktop, photo-realistic virtual environments. 16th Annual Colloquium of the Spatial Information Research Centre (SIRC). Dunedin, New Zealand, 2004.
[5] Yoon, S.Y., Laffey, J., Oh, H. Understanding usability and user experience of web-based 3D graphics technology. International Journal of Human-Computer Interaction 24 (3), 2008, pp. 288-306.
[6] Marsh, T. & Wright, P. Co-operative evaluation of a desktop virtual reality system. In: Harrison, M., Smith, S. (Eds.), Proceedings of the Workshop on User Centered Design and Implementation of Virtual Environments, University of York, UK, 1999, pp. 99-108.
[7] Rosli, D.I., Chen, C.J., Teh, C.S. Interface design usability testing of Virtual Reality (VR)-based learning environment. Proceedings of the 5th International Conference on Information Technology and Applications, ICITA 2008, Queensland, Australia.
[8] Costalli, F., Marucci, L., Mori, G., Partenò, F. Design criteria for usable web accessible virtual environments. In: Proceedings of ICHIM'01, Milan, Italy: Politecnico di Milano and Archives & Museum Informatics, 2001, pp. 413-426.
[9] Rezazadeh, I. M., M. Firoozabadi, M., & Wang, X. (2011). Evaluating the usability of virtual environment by employing affective measures. *Intelligent Systems, Control and Automation: Science and Engineering, 1010,* 95-109.
[10] Downey, L.L.. Group usability testing: Evolution in usability techniques. Journal of Usability Studies Vol. 2 No.3, 2007, pp.133-144.
[11] Smith, S.P., Hart, J. Evaluating distributed cognitive resources for wayfinding in a desktop virtual environment. In: Kitamura, Y., Bowman, D., Fröhlich, B., Stürzlinger, W. (Eds), IEEE Symposium on 3D User Interfaces 2006, pp. 3-10.
[12] Seffah, A., Donyaee, M. Kline. R.B., Padda, H.K.,. Usability measurement and metrics: A consolidated model. Software Qual J 14, 2006, pp. 159-178.
[13] Dumas, J.S., Redish, J. C.,. A practical guide to usability testing, Ablex Publishing, Norwood, NJ, 1993.
[14] Nielsen, J. Severity ratings for usability problems., 1995. Available from: <www.useit.com/papers/heuristic/ severityrating.html>

[15] Law, L., Hvannberg, E.T. Complementarity and convergence of heuristic evaluation and usability test: A case study of universal brokerage platform. In Bertelsen, O.W., Bodker, S., Kuutti, K. (Eds.), Proceedings of NordiCHI 2002. ACM, NY, 2002, pp. 71-80.

[16] Doubleday, A., Ryan, M., Springett, M., Sutcliffe, A.,. A comparison of usability techniques for evaluating design. In: Proceedings of the conference on Designing interactive systems: processes, practices, methods, and techniques, Amsterdam, The Netherlands, 1997, pp.101-110.

[17] Jeffries, R., Miller, J.R., Wharton, C., Uyeda, K.M. User interface evaluation in the real world: A comparison of four techniques. Proceedings of the ACM CHI'91 Conference on Human Factors in Computer Systems, 1991, pp. 119-124.

[18] Faulkner L. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. Behavior Research Methods, Instruments, & Computers 35 (3), 2003, pp. 379-383.